**ABSTRACT**

A means for minimizing time for a system/device initial program load (IPL) in a system that will not support instruction prefetching when executing IPL code out of non-volatile memory. The IPL code is organized into a first portion and second portion. The first portion is executed from the non-volatile memory device to configure system memory; the first portion also provides initial control of cache inhibit and cache enable by way of software control. The cache-enabling code is strategically located at a memory page boundary such that the system hardware will disable instruction prefetching in an adjoining page just past this cache enabling software code. After the first portion of IPL code configures system memory, the second portion is copied into memory through the L2 cache and executed from memory with cache enabled to allow speculative instruction prefetching.